NanoNet: Real-Time Polyp Segmentation in Endoscopy

Debesh Jha*†, Nikhil Kumar Tomar*, Sharib Ali§, ¶ Michael A. Riegler*,
Håvard D. Johansen†, Dag Johansen†, Pål Halvorsen*‡

*SimulaMet, Norway †UiT The Arctic University of Norway, Norway ‡Oslo Metropolitan University, Norway
§Department of Engineering Science, Big Data Institute, University of Oxford, Oxford, UK
¶Oxford NIHR Biomedical Research Centre, Oxford, UK
Email: debesh@simula.no

***Abstract*—Deep learning in gastrointestinal endoscopy can assist to improve clinical performance and help to obtain a more accurate patient assessment. To this extent, semantic segmentation methods that can perform automated real-time delineation of a region-of-interest, e.g., boundary identification of cancer or precancerous lesions, can benefit both diagnosis and interventions. However, accurate and real-time segmentation of endoscopic images is extremely challenging due to its high operator dependence and high-definition image quality. To utilize automated methods in clinical settings, it is crucial to design lightweight models with low latency such that they can be integrated with low-end endoscope hardware devices. In this work, we propose a novel architecture for the segmentation of endoscopic images *NanoNet*. Our proposed architecture allows a real-time performance and has higher segmentation accuracy compared to other more complex ones. We use wireless capsule endoscopy and standard colonoscopy datasets with polyps, and a dataset consisting of endoscopy biopsies and surgical instruments, to demonstrate the effectiveness of our model. Our experiments demonstrate the increased performance of our method in terms of a trade-off between model complexity, speed, model parameters, and metric performances. Moreover, the resulting models´ size is relatively tiny, with only nearly 36,000 parameters compared to traditional deep learning approaches having millions of parameters.**

***Index Terms*—Medical image analysis, deep learning, segmentation, colonoscopy, wireless capsule endoscopy, tool segmentation**

I. Introduction

Gastrointestinal (GI) endoscopy is widely used technique to diagnose and treat anomalies in the upper (esophagus, stomach, and duodenum) and the lower (large intestine and anus) GI tract. Among the other GI tract organs, colorectal cancer (CRC) has the highest cancer incidences and mortality rate [1]. There are several CRC screening options. Theses are usually divided into two categories, namely, invasive (visual examination-based test) and non-invasive based tests (stool, blood, and radiological test). *Colonoscopy* is the invasive examination procedure. It is used to observe and remove the abnormalities (such as polyps) inside the large intestine (also known as the colon) and rectum. It is considered the gold standard for the examination of colorectal cancer with both high sensitivity and specificity. The others are *Sigmoidscopy*, *Computed Tomography(CT) Colonoscopy*, *Narrow-Band Imaging (NBI)*, *Fecal Occult Blood Test (FOBT) and Fecal Immunochemical Test (FIT)*. Video Capsule Endoscopy (VCE) is a technology for capturing the video inside the GI tract. It has evolved as an important tool for screening small bowel diseases [2].

Deep Learning (DL) methods have made a significant breakthrough in several medical domain such as lung cancer detection [3], diabetic retinopathy progression [4], obstructive hypertrophic cardiomyopathy detection [5]. It has provided new opportunities to solve the challenge such as bleeding, light over/underexposure, smoke, and reflections [6]. However, DL needs a large annotated dataset for the implementation of methods. It is difficult to obtain a labeled medical dataset. First, it needs collaborations with the hospitals. For data collection, the doctors require approval from the hospital and patient consent. They need to set protocols for the collection, and the collected data must be anonymized and cleaned with the help of computer scientists. The expert must label raw data. After labeling, the annotations must be done depending upon the need the task. The whole process requires an enormous amount of expert time and is costly. Additionally, it is an operator-dependent process. The quality of the data labeling and annotation depends on the expertise of the clinicians. Therefore, it is challenging to curate a larger dataset.

One way of solving the dataset issue is to create synthetic images using Genearative Adversarial Network (GAN). However, generated synthetic images may not always capture all the properties and characteristics of real endoscopic images. Consequently, the model may only learn to predict the properties from the synthetic images and may not perform well on a real endoscopic dataset. Another solution could be domain adaptation from a similar endoscopic dataset. However, we lack a large publicly available labeled endoscopic datasets. Thus, an excellent way to solve the semantic segmentation task is to reuse ImageNet pre-trained encoders in the segmentation model [7]. The predicted masks from the algorithm can provide reliable information to the endoscopic model.

A lightweight Convolutional Neural Network (CNN) model can be essential for the development of real-time and efficient semantic segmentation methods. Usually, lightweight models are computationally efficient and require less memory. A smaller number of parameters make the network less redundant. Lightweight CNN models are mainly being deployed in mobile applications [8]. A lightweight model can play a crucial role from a system perspective with a limited resource

constraint for real-time prediction in clinics. We propose a novel architecture, NanoNet, optimized for faster inference and high accuracy. An extremely lightweight model with very few trainable parameters, faster inference, and higher performance would require less memory footprint to be incorporated with any devices. Therefore, we put forward this approach to address the challenges in endoscopy.

The main contributions of this work are as following:

1) We proposed a novel architecture, named NanoNet, to segment endoscopic images in real-time with high accuracy. The proposed architecture is exceptionally lightweight, the model size is smaller, and it requires less computational cost.
2) VCE datasets are difficult to obtain with pixel-wise annotations. In this context, we have annotated 64 polyps from the "polyp" class of the Kvasir-Capsule dataset with the help of an expert gastroenterologist and provide the benchmark.
3) NanoNet achieves promising performance on the KvasirCapsule-SEG, Kvasir-SEG [9], Medico 2020 [10], EndoTect 2020 [11], and Kvasir-Instrument [12] datasets. All experiments show the state-of-the-art (SOTA) trade-off in terms of parameter uses (size), speed, computation, and performance metrics.
4) The model can be integrated with mobile and embedded devices because of fewer parameters used in the network.

II. Related work

A. Semantic segmentation of endoscopic images

Semantic segmentation of endoscopic images has been a well-established topic in medical image segmentation. Earlier work mostly relied on the handcrafted descriptors for feature learning [13], [14]. The handcrafted features such as color, shape, texture, and edges were extracted and fed to the Machine Learning (ML) classifier, which separates lesions from the background. However, the traditional ML methods based on handcrafted features suffer from low performance [15]. The recent works on endoscopic image segmentation mostly relied on Deep Neural Network (DNN) [16]–[22].

With the DNN methods, there is progress in the performance for segmenting endoscopic images (for example, polyps). However, the network architectures are often complex that requires high-end GPU for training, and is computationally expensive [21], [23], [24]. Additionally, real-time lesion segmentation has been ignored often. Although there is some recent initiation for the real-time detection of endoscopic images, they have mostly used private datasets [25]–[27] for the experimentation. It is difficult to compare the new methods on these datasets and extend the benchmark. Therefore, there is a need for the benchmark on publicly available datasets to minimize the research gap towards building a clinically relevant model.

B. Lightweight model

There are few works already proposed that uses the concept of a lightweight model. Ni et al. [28] presented a novel bilinear attention network-based approach with an adaptive receptive field for the segmentation of surgical instruments. Wang et al. [29] proposed a lightweight encoder-decoder network (LEDNet), an encoder-decoder network uses ResNet50 in the encoder block and attention pyramidal network in the decoder block. Beheshti et al [30] proposed SqueezeNet. The architecture of the SqueezeNet is inspired by UNet [31]. The proposed model obtained $12\times$ reduction in model size and showed efficient performance in multiplication accumulation (mac) and memory uses.

From the above-related work, we can see that there is a need for a real-time polyp segmentation method. A real-time polyp segmentation method can be achieved by building a lightweight network architecture which can be build by designing an efficient network with blocks that require fewer parameters. A lower number of network parameters will reduce the network complexity, leading to real-time or faster inference. In this respect, we propose NanoNet, which uses a lightweight pre-trained network MobileNetV2 [32], and simple convolutional blocks such as residual block and squeeze and excite block.

III. Network architecture

The architecture of the NanoNet follows an encoder-decoder approach as shown in Figure 1. As depicted in Figure 1, the network architecture uses a pre-trained model as an encoder, followed by the three decoder blocks. Using pre-trained ImageNet [33] models for transfer learning has become the best choice for many CNN architecture [7], [24]. It helps the model converge much faster and achieve high performance compared to the non-pre-trained model. The proposed architecture uses a MobileNetV2 model pre-trained on the ImageNet dataset as the encoder. The decoder is built using a modified version of the residual block, which was initially introduced by He et al. [34]. The encoder is used to capture the required contextual information from the input, whereas the decoder is used to generate the final output by using the contextual information extracted by the encoder.

A. MobileNetV2

The MobileNetV2 is an architecture that is primarily designed for mobile and embedded devices. The architecture performed well on a variety of different datasets while maintaining high accuracy, despite having fewer parameters. The architecture of MobileNetV2 is based on the architecture of MobileNetV1, which uses depth-wise separable convolutions as the main building block. A depth-wise separable convolution consists of depth-wise convolution followed by a point-wise convolution. The MobileNetV2 introduces two main ideas: Inverted residual block and linear bottleneck block [32].

The inverted residual block is based on the bottleneck residual block as described in the [34], which consists of three standard convolutions, which are 1×1, 3×3, and 1×1. Every convolution layer is followed by a Rectified Linear Unit (ReLU) non-linearity. In the first 1×1 standard convolution, the number of feature channels are reduced, and

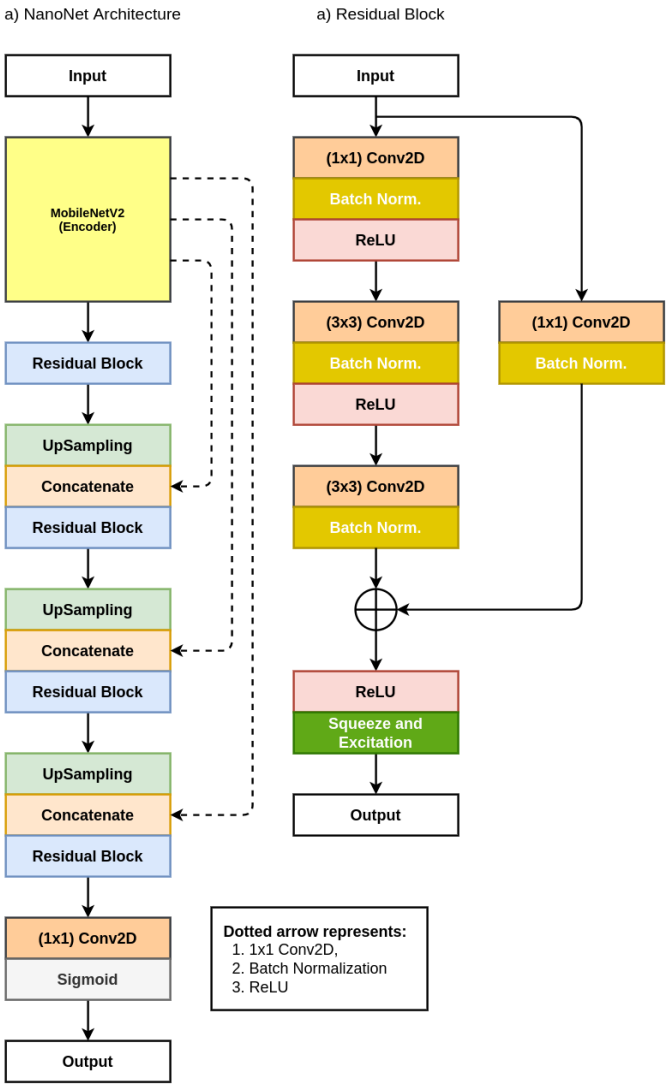

Fig. 1: Overview of the proposed NanoNet architecture

in the last 1×1 standard convolution, the number of feature channels are expanded. After that, an element-wise addition with the identity mapping is performed. The inverted residual block also has three convolution layers: a 1×1 standard convolution, a 3×3 depth-wise convolution, and a 1×1 standard convolution. Every convolution has a ReLU activation function. Here, the exact opposite of the bottleneck residual block is performed. The first 1×1 standard convolution expands the number of feature channels, and the last 1×1 standard convolution reduces the number of feature channels. Due to this opposite functionality, it is referred to as an inverted residual block. The linear bottleneck block is the same as the inverted residual block, except the last 1×1 standard convolution has a linear activation before an element-wise addition is performed with the identity mapping.

B. Modified Residual Block

The original residual block uses two 3×3 standard convolutions, where the first convolution is followed by a batch-normalization and a ReLU activation function. After that, the second convolution is followed only by a batch-normalization. An element-wise addition is performed between the output of the batch-normalization and the identity mapping, followed by another ReLU activation function. An identity mapping consists of a 1×1 standard convolution and a batch-normalization over the original input.

We have modified the residual block for our network. The modified residual block starts with a 1×1 convolution followed by a 3×3 convolution. In both of these convolutions, we reduce the number of filters by $\frac{1}{4}$, which are then followed by the batch normalization and the ReLU activation function. We have a 3×3 convolution with batch normalization. Now, we perform an element-wise addition with the identity mapping. Finally, we apply a ReLU activation function followed by the squeeze and excitation block. The squeeze and excitation block improves the quality of feature maps by increasing their sensitivity towards essential features.

C. The NanoNet architecture

Figure 1 shows the block diagram of the NanoNet architecture. The NanoNet architecture starts with a pre-trained MobileNetV2 as an encoder followed by a decoder. There is a modified residual block between the encoder and the decoder, which acts like a bridge that connects the encoder and the decoder. In the first step, we feed the image data into the pre-trained encoder. The pre-trained encoder starts with a standard convolution with 32 feature channels, followed by the bottleneck layer with ReLU6 as the activation function. All the convolution operation uses a standard 3×3 kernel size. The entire encoder network progressively downsamples the feature maps by using strided convolution and slowly increases the number of feature channels alternatively.

The output from the pre-trained encoder passes through the modified residual block, which is fed to the decoder. Every step in the decoder uses a bilinear upsampling to increase the spatial dimension (height and width) of the input feature maps. After that, it is concatenated with the appropriate feature maps from the pre-trained encoder using the skip connections. These skip connections pass information that may be lost sometimes between the layers and are used to improve the quality of the feature maps. These concatenated feature maps are passed through the modified residual block, which further increases the generalization capacity of the decoder. After the feature maps pass through all the three decoder block, the output of the last decoder block is fed to a 1×1 convolution with a number of classes as the feature channels, which is followed by the sigmoid activation if it is a binary segmentation task, else we use softmax activation function.

We have demonstrated three different NanoNet architectures: NanoNet-A, NanoNet-B, and NanoNet-C. Each architecture consists of different feature channels in its decoder block. NanoNet-A consists of 32, 64 and 128 feature channels. In NanoNet-B, the number of feature channels reduces to 32, 64, and 96. In NanoNet-C, these feature channels are further reduced to 16, 24, and 32. The reduction in the number of feature channels leads to less trainable parameters, which reduces the model complexity leading to a light-weight network.

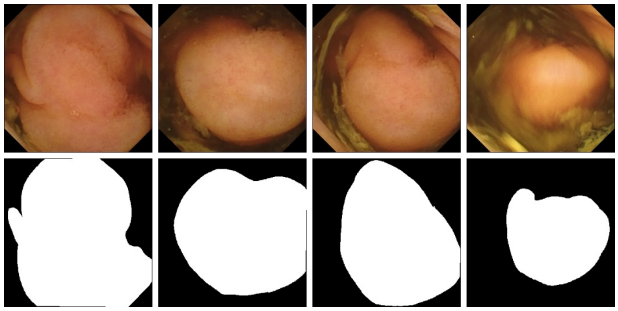

Fig. 2: Polyps and corresponding masks from KvasirCapsule-SEG

IV. Experimental setup

In this section, we will describe about the dataset, experimental setup, and data augmentation techniques used.

TABLE I: Publicly available medical imaging datasets used in our experiments.

Dataset	# of Images	Imaging Type
Kvasir-SEG [9]	1000	Colonoscopy
Kvasir-Instrument [35]	590	Colonoscopy
Medico Challenge Dataset [10]	160	Colonoscopy
Endotect Challenge Dataset [11]	200	Colonoscopy
KvasirCapsule-SEG	60	WCE

A. Datasets

To address the incidence of CRC, we have selected the polyp class of the Kvasir-Capsule dataset [36] and annotated it with the help of gastroenterologists. The Kvasir-Capsule is an open-access dataset that contains 13 classes of anomalies and findings. It only includes 64 polyp frames out of 44,228 medically verified video capsule frames present in the Kvasir-Capsule. We have annotated the polyp class of Kvasir-Capsule and generated corresponding ground truth masks. Examples of polyps and their corresponding masks from KvasirCapsule-SEG can be found in Figure 2. Furthermore, we also provide bounding box information to be used for video capsule endoscopy detection and localization tasks. The Kvasir-Capsule can be downloaded easily from here [1]. Additionally, Table I shows the detail information about the other open imaging dataset used in our experiments. The detail explanation of the other datasets can be found in [9]–[12].

For the evaluation of our model, we have chosen standard computer vision metrics such as Dice Coefficient (DSC), mean Intersection over Union (mIoU), Precision, Recall, Specificity, Accuracy, and Frame-per-second (FPS). More explanation of these metrics can be found in [9]–[12].

B. Implementation details

We have implemented the NanoNet using Keras[2] with TensorFlow [39] as backend. The experiments were run on the Experimental Infrastructure for Exploration of Exascale Computing (eX3), NVIDIA DGX-2 machine. As the model has very few low trainable parameters, we have set a batch size of 16. We have resized the dataset images to 256×256 pixels for better utilization of the GPU, and it also helps to reduce the training time. The model is trained on 200 epochs with Nadam optimizer [40] and dice coefficient as the loss function. The learning rate for the optimizer is set to $1e^{-4}$. We prefer to choose a low learning rate to update the parameters slowly and carefully. The learning rate is reduced by a factor of 0.1 when the validation loss does not decrease in 10 consecutive epochs. It helps to improve model performance. Additionally, we have used an early stopping mechanism to prevent overfitting.

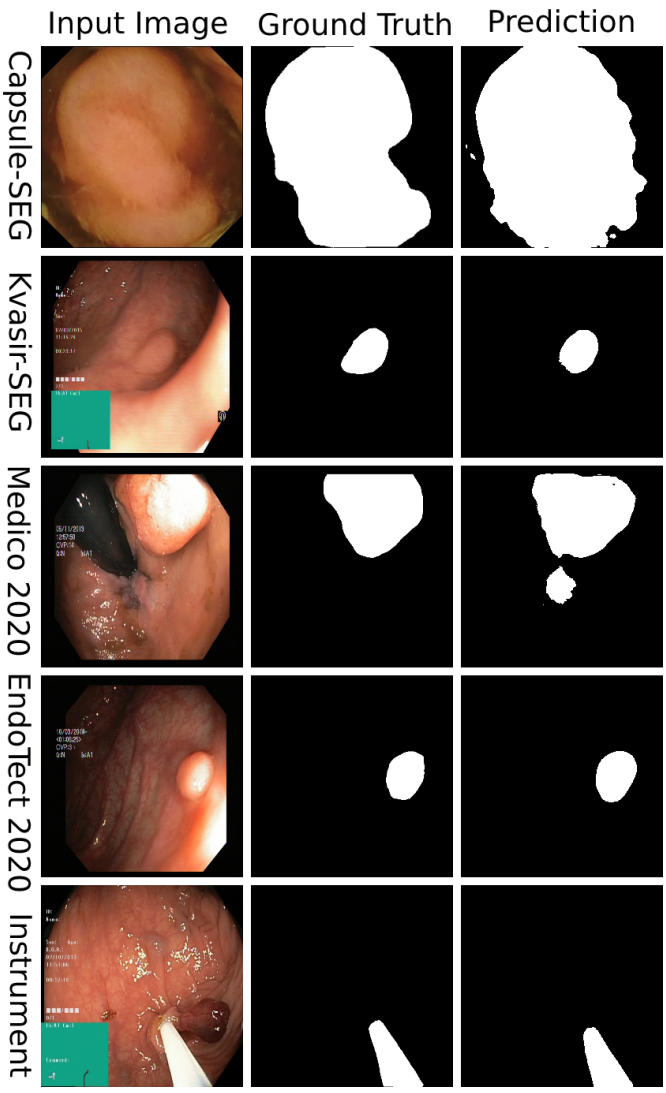

Fig. 3: Qualitative results of NanoNet-A on five different datasets.

C. Data augmentation

We use data-augmentation on the training set to increase diversity and to improve the generalization of our model. Data augmentation techniques such as random cropping, random rotation, horizontal flipping, vertical flipping, grid distortion,

[1] https://osf.io/dv2ag/
[2] https://keras.io/

TABLE II: Performance evaluation of the proposed networks and recent SOTA methods on KvasirCapsule-SEG

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [37]	8,227,393	**0.9206**	0.8687	0.9444	0.9086	0.9328	**0.9352**	16.00
ResUNet++ (ISM'19) [23]	4,070,385	0.9150	0.8551	0.9319	0.9111	0.9282	0.9282	15.36
NanoNet-A (Ours)	235,425	0.9205	**0.8719**	**0.9681**	0.8958	**0.9440**	0.9340	28.49
NanoNet-B (Ours)	132,049	0.9137	0.8596	0.9076	**0.9276**	0.9094	0.9390	29.14
NanoNet-C (Ours)	36,561	0.9101	0.8577	0.9510	0.8941	0.9291	0.9246	**32.31**

TABLE III: Performance evaluation of the proposed networks and recent SOTA methods on Kvasir-SEG [9]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [37]	8,227,393	0.7203	0.6106	0.7602	0.7624	0.7327	0.9251	17.72
ResUNet++ (ISM'19) [23]	4,070,385	0.7310	0.6363	0.7925	0.7932	0.7478	0.9223	19.79
NanoNet-A (Ours)	235,425	**0.8227**	**0.7282**	**0.8588**	**0.8367**	**0.8354**	**0.9456**	26.13
NanoNet-B (Ours)	132,049	0.7860	0.6799	0.8392	0.8004	0.8067	0.9365	29.73
NanoNet-C (Ours)	36,561	0.7494	0.6360	0.8081	0.7738	0.7719	0.9290	**32.17**

TABLE IV: Performance evaluation of the proposed networks and recent SOTA methods on the Medico 2020 dataset [10]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [37]	8,227,393	0.6846	0.5599	0.7235	0.7236	0.6961	**0.9231**	18.54
ResUNet++ (ISM'19) [23]	4,070,385	0.6925	0.5849	0.8249	0.6840	0.7434	0.8995	19.47
NanoNet-A (Ours)	235,425	0.7364	**0.6319**	**0.8566**	0.7310	**0.7804**	0.9166	28.07
NanoNet-B (Ours)	132,049	**0.7378**	0.6247	0.8283	**0.7373**	0.7685	0.9223	29.04
NanoNet-C (Ours)	36,651	0.7070	0.5866	0.8095	0.7089	0.7432	0.9148	**32.66**

TABLE V: Performance evaluation of the proposed networks and recent SOTA methods on the Endotect 2020 dataset [11]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [38]	7,771,873	0.6640	0.5408	0.7510	0.6841	0.6943	0.9075	26.55
ResUNet++ (ISM'19) [23]	4,070,385	0.6940	0.5838	**0.8797**	0.6591	0.7597	0.8841	18.58
NanoNet-A (Ours)	235,425	**0.7508**	**0.6466**	0.8238	**0.7744**	**0.7773**	**0.9255**	27.19
NanoNet-B (Ours)	132,049	0.7362	0.6238	0.8109	0.7532	0.7646	0.9252	29.91
NanoNet-C (Ours)	36,651	0.7001	0.5792	0.8000	0.7159	0.7380	0.9091	**32.98**

TABLE VI: Performance evaluation of the proposed networks and recent SOTA methods on Kvasir-Instrument [12]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
UNet (Baseline) [38]	-	0.9158	0.8578	**0.9487**	0.8998	**0.9320**	0.9864	20.46
DoubleUNet (Baseline) [24]	-	0.9038	0.8430	0.9275	0.8966	0.9147	0.9838	10.00
ResUNet++ (ISM'19) [23]	4,070,385	0.9140	0.8635	0.9103	0.9348	0.9140	0.9866	17.87
NanoNet-A (Ours)	235,425	0.9251	0.8768	0.9142	**0.9540**	0.9251	**0.9887**	28.00
NanoNet-B (Ours)	132,049	**0.9284**	**0.8790**	0.9205	0.9482	0.9284	0.9875	29.82
NanoNet-C (Ours)	36,561	0.9139	0.8600	0.9037	0.9452	0.9139	0.9863	**32.18**

and many more are used. We have used an offline data augmentation technique. The validation and testing set is not augmented and is directly resized into 256×256.

V. Result and Discussion

In this section, we provide the experimental results for the segmentation task of the endoscopic image dataset. For the evaluation, we have used performance metrics such as DSC and mIoU, and FPS as the main evaluation metrics. We also calculate recall, precision, F2, and overall accuracy to support a complete set of metrics. Table II, Table III, Table IV, Table V, and Table VI show the test results of the NanoNet models using different parameters. The results are compared with the recent SOTA computer vision methods. The qualitative results are displayed in Figure 3. The first, second, and third columns show the image, ground-truth, and prediction masks, respectively. Similarly, the name of the dataset is provided on the left side. One example image for each dataset is shown. The qualitative results with diversified classes of medical datasets show that NanoNet can produce accurate segmentation results with different types of lesions (polyps) and therapeutic tools. For example, the examples images and the prediction show that NanoNet produces good segmentation masks for large, medium, and small polyps (see Figure 3). From the qualitative results, we can see that NanoNet produces good results with small-sized polyps but produces over-segmentation for the large-sized lesions. For future work one could create a specific dataset consisting of a set of small and large-sized polyps to explore this further.

The quantitative results show that NanoNet consistently outperforms or performs nearly equal to its competitor in terms of performance. The quantitative results also show that NanoNet is capable of producing real-time segmentation (i.e., produces at least 30 FPS for each dataset present in the Tables. This is one of the major contributions of the work. The other strength of the work lies in the parameter uses. From Table II, we can see the best performing NanoNet (i.e., NanoNet-A) uses nearly 35 times less as parameter as ResUNet [37]. Similarly, NanoNet-C uses 225 times less parameter as compared to

that of ResUNet and also produces better DSC, mIoU and FPS with the Kvasir-SEG. Thus from both evaluation metrics and qualitative results, the improvement is overwhelming. Thus, the proposed NanoNet architecture is simple, compact, and provides a robust solution for real-time applications, as it produces satisfactory performance despite having fewer parameters.

VI. Conclusion

In this paper, we proposed a novel lightweight architecture for real-time endoscopic image segmentation. The proposed NanoNet architecture utilizes a pre-trained MobileNetV2 model and a modified residual block. The depth-wise separable convolution is the main building block of the network and allows the model to achieve high performance with minuscule trainable parameters. The experimental results on varied endoscopy datasets demonstrate the strength of our model compared to state-of-the-art models with respect to their speed and performance. The presented model has the potential to enable easier role-out of deep learning models in clinical systems due to fewer parameters, competitive accuracy, and low-latency. In addition, the model does not require any sort of initialization, post-processing or temporal regularization, which can be considered as another strength of this work. In the future, we will design an encoder lighter than the currently used pre-trained MobilNetV2. Moreover, we aspire to utilize the currently built segmentation module in the clinic and study the efficacy of our designed model.

References

[1] H. Sung *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, 2021.

[2] A. Kornbluth, P. Legnani, and B. S. Lewis, "Video capsule endoscopy in inflammatory bowel disease: past, present, and future," *Inflammatory Bowel Diseases*, vol. 10, no. 3, pp. 278–285, 2004.

[3] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[4] F. Arcadu *et al.*, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[5] E. M. Green *et al.*, "Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–4, 2019.

[6] S. Bodenstedt *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv preprint arXiv:1805.02475*, 2018.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[8] Y.-D. Kim *et al.*, "Compression of deep convolutional neural networks for fast and low power mobile applications," *arXiv preprint arXiv:1511.06530*, 2015.

[9] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. of International Conference on Multimedia Modeling (MMM)*, 2020, pp. 451–462.

[10] D. Jha and Others, "Medico multimedia task at mediaeval 2020: Automatic polyp segmentation," in *CEUR Proceedings of MediaEval Workshop*, 2020.

[11] S. A. Hicks *et al.*, "The endotect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy," in *Proceedings of ICPR 2020 Workshops and Challenges*, 2020.

[12] D. Jha *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *Proc. of Multimedia Modeling (MMM)*, 2021.

[13] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 3, pp. 141–152, 2003.

[14] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009*, 2009, pp. 346–350.

[15] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[16] M. Min *et al.*, "Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[17] S. Ali *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[18] S. Ali, M. Dmitrieva *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical Image Analysis*, p. 102002, 2021.

[19] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[20] Y. Guo, J. Bernal, and B. J Matuszewski, "Polyp segmentation with fully convolutional deep neural networks—extended evaluation study," *Journal of Imaging*, vol. 6, no. 7, p. 69, 2020.

[21] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020, pp. 263–273.

[22] D. Jha *et al.*, "A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation." *IEEE Journal of Biomedical and Health Informatics*.

[23] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *Proc. of IEEE International Symposium on Multimedia (ISM)*, 2019, pp. 225–2255.

[24] D. Jha, , M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *Proc. of International Conference on Multimedia Modeling (MMM)*, 2020, pp. 451–462.

[25] J. Y. o. Lee, "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[26] M. Yamada *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[27] C. C. Poon *et al.*, "Ai-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[28] Z.-L. Ni *et al.*, "Barnet: Bilinear attention network with adaptive receptive field for surgical instrument segmentation," *arXiv preprint arXiv:2001.07093*, 2020.

[29] Y. Wang *et al.*, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1860–1864.

[30] N. Beheshti and L. Johnsson, "Squeeze u-net: A memory and energy efficient image segmentation network," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 364–365.

[31] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[33] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, 2009, pp. 248–255.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[35] J. Bernal *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
[36] P. H. Smedsrud *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Springer Nature Scientific Data*, 2021.
[37] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015, pp. 234–241.
[39] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. of USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
[40] T. Dozat, "Incorporating nesterov momentum into adam," in *Proc. of International Conference on Learning Representations*, 2016.